Les Learning Analytics vus par la sociologie

Paola Tubaro

Les *Learning Analytics* vus par la sociologie[1]

Paola Tubaro, CNRS[2]

L'intérêt croissant pour les *Learning Analytics* (LA) reflète une transformation dans la pratique et la science de l'éducation, parallèle aux bouleversements qu'ont apportés les données numériques massives (« *big data* ») dans d'autres domaines, allant des sciences à la finance, l'industrie et la santé. C'est dans cet effort de mise en perspective et de comparaison que s'inscrit ma contribution à ce débat dans la revue DMS, n'étant pas spécialiste des LA en particulier ni des sciences de l'éducation en général, mais ayant animé la réflexion dans une discipline voisine, la sociologie, notamment avec la coordination d'un numéro spécial de la *Revue Française de Sociologie* en 2018 (Bastin et Tubaro, 2018a).

La promesse de fond des LA est la possibilité, par l'accès à des grandes masses de données, de mieux saisir les processus d'apprentissage et potentiellement, de les améliorer. Ces données sont les traces que laissent les élèves dans un espace numérique d'apprentissage où ils réalisent des activités, comme la réponse à des quiz ou questions, la publication de messages dans des forums de discussion, la prise de parole dans des séances de travail en ligne etc. Pour l'essentiel, les LA consistent à extraire ces informations, les agréger en des indicateurs interprétables, et les mettre à disposition. Le bénéficiaire peut être l'individu apprenant qui souhaite surveiller son progrès, et/ou l'enseignant qui vise à suivre l'ensemble de la classe et détecter d'éventuels axes d'amélioration[3]. Bien évidemment il s'agit aussi d'un outil formidable au service de la recherche en sciences de l'éducation, lui fournissant une information très détaillée à l'échelle de l'individu, disponible en continu sur des périodes assez longues, et potentiellement étendue à de grands nombres d'individus.

Bien sûr il y a plusieurs interprétations de ces principes basiques, et un grand mérite du débat qui a eu lieu tout au long de cette année dans la revue DMS (Peraya, 2019a), a été de montrer cette variété en présentant des expérimentations en cours dans différents établissements, surtout universitaires. Les contributions des auteurs montrent bien, par exemple, que les utilisations des LA pour la recherche et pour l'enseignement suivent des logiques différentes et parfois difficiles à concilier (Peraya, 2019c), qu'une démarche éthique est nécessaire pour que l'outil puisse véritablement rendre service aux apprenants (Peraya, 2019b), que l'application de celui-ci diffère fortement dans le contexte d'un petit Environnement Informatique pour l'Apprentissage Humain (EIAH) par rapport à un MOOC, la quantité de données produites dans le second cas étant bien plus importante.

À la lumière de ces réflexions, il me semble utile de revenir sur des enjeux que ces données, et leur usage dans le cadre des LA, font apparaître, et qui sont communs à d'autres contextes d'application. Je ne reviendrai pas sur la définition des *big data* (Bastin et Tubaro, 2018b), ni sur celle des LA (Peraya, 2019a), mais je m'efforcerai de dresser un bilan, en mettant en relation l'état de l'art en sociologie, avec les éléments principaux qui ressortent du débat dans la DMS. Dans ce qui suit, je reviendrai d'abord sur les prédictions de « révolution » souvent associées à l'essor des *big data*, pour montrer qu'elles étaient bien en décalage par rapport aux faits, tels qu'on peut les observer aujourd'hui. Ensuite, je montrerai que le

[1] Publié dans *Distances et médiations des savoirs* [En ligne], 28 | 2019, mis en ligne le 16 décembre 2019. URL : http://journals.openedition.org/dms/4608

[2] Laboratoire de Recherche en Informatique (LRI), Bâtiment 660 Université Paris Sud, 91405 Orsay Cedex, France. Email : paola.tubaro@lri.fr

[3] Dans ce cas, le respect de la vie privée des usagers peut imposer que l'enseignant reçoive seulement une version anonymisée et agrégée des données.

potentiel de ces données et techniques est loin d'être épuisé, mais que sa réalisation, loin de s'appuyer sur des automatismes, nécessite une réorganisation adaptée des éléments humains qui l'entourent.

La révolution des big data n'a jamais eu lieu… ou presque

Qu'avons-nous appris, jusqu'aujourd'hui, de la révolution des *big data* (à laquelle se rattachent les lA) annoncée il y a désormais plus de dix ans (Savage et Burrows, 2007) ? Mon premier constat, après avoir exploré le champ en sociologie (Bastin et Tubaro, 2018b) et en sciences de l'éducation grâce au débat de DMS, est que contrairement aux prédictions des débuts, ces données massives n'ont pas remplacé les autres modes de production de la connaissance, mais s'y sont plutôt rajoutés, en complément à ceux-ci ; la révolution n'a pas eu lieu, ou du moins, elle ne s'est pas déroulée de la manière initialement attendue, et n'a amené ni la « fin de la théorie » (Anderson, 2008), ni la disparition d'autres modes de collecte de l'information et de construction de la connaissance, comme les questionnaires, les entretiens et les observations traditionnellement réalisés en sociologie (et dans d'autres disciplines au sein des sciences sociales). Pourquoi ?

Une première réponse à cette question est que les *big data* ne tracent pas tout le monde. La sociologie insiste sur les inégalités d'accès aux outils numériques, et les voix qui s'expriment (par exemple) sur les médias sociaux sont celles des couches sociales les plus favorisées (Hargittai, 2018). Même parmi les jeunes en âge scolaire, il existe des variations importantes en termes d'accès, d'usages et de compétences, avec des répercussions parfois fortes sur leur performance académique ainsi que, plus tard, sur leurs perspectives d'emploi (Robinson *et al*., 2015). Des environnements d'apprentissage en ligne, qu'il s'agisse de MOOC ou d'espaces plus petits (utilisés pour la formation à distance ou même à côté d'une classe présentielle), peuvent alors attirer des publics spécifiques : s'il en est ainsi, l'analyse des traces que laissent ceux-ci n'est pas représentative des processus d'apprentissage dans la population, plus hétérogène, que des programmes éducatifs de même type ou niveau pourraient cibler. En se servant de données produites par les usagers de ces espaces numériques, alors, on risque d'aboutir à des conclusions biaisées. Si des projets formatifs sont ensuite construits sur cette base, sans tenir compte des besoins et des comportements des plus désavantagés, ils finiront paradoxalement par creuser davantage les inégalités.

Une autre insuffisance des *big data* est qu'il y a bien souvent un arbitrage à faire entre la taille et la richesse des bases de données. Les traces horodatées de comportements que l'on peut extraire d'environnements numériques sont précieuses pour évaluer, par exemple, le progrès d'un élève, ou d'un groupe d'élèves, du début à la fin d'une formation. Sur cette base, il est possible d'établir de grandes tendances, et c'est ce qui fait l'intérêt et la force des LA ; il est pourtant plus difficile de les expliquer. La complication vient du fait que, malgré leur abondance apparente, ces données sont « souvent parcimonieuses pour chaque apprenant » (Boyer, 2019). De même, une autre auteure affirme que « plus les données sont massives, plus elles sont factuelles et donc moins elles possèdent d'épaisseur », cette notion indiquant, selon sa définition, la disponibilité d'informations sur le contexte, essentielles pour interpréter une donnée (Peraya et Luengo, 2019). Par exemple, sans information sur le statut socioéconomique des élèves, il est impossible d'établir dans quelle mesure des inégalités sociales héritées du monde non-numérique se répercutent sur leur engagement en ligne. On est loin de la richesse des bases de données classiques, comme les enquêtes menées depuis longtemps en sociologie, construites précisément pour les besoins de l'analyse et par là plus aptes à aider l'interprétation, malgré leur petite taille et leur caractère déclaratif.

Pour « enrichir » les *big data* afin d'en extraire davantage de connaissance, il est souvent préconisé d'apparier différentes sources de données. Ce n'est pas par hasard que vers la fin des années 2000, les

grandes multinationales du numérique en pleine croissance, promouvaient avec enthousiasme l'idée que la protection de la vie privée aurait perdu son intérêt, les usagers du web étant séduits par la possibilité de se dévoiler autant que possible. Aujourd'hui, les attitudes ont pourtant changé et l'attention à la vie privée s'est imposée, dans les pratiques des individus tout comme dans la règlementation – notamment avec le RGPD[4] en Europe (Tubaro, Casilli et Sarabi, 2014 ; Tubaro, 2019). La collecte de traces de comportements en ligne, y compris dans des environnements construits pour l'apprentissage, doit donc suivre des règles, légales et éthiques, de plus en plus strictes. Il est donc parfois nécessaire de renoncer à collecter certaines données, ou à apparier des bases de données différentes, si les conditions prévues par la loi et la déontologie ne sont pas remplies (par exemple, en raison du manque de consentement de la personne concernée), même si la technologie permettrait de le faire. La prise de conscience des dérapages possibles amène les répondants à demander davantage de garanties.

Une autre limite des *big data* est que malgré la puissance des algorithmes qui y sont appliqués, les données brutes sont difficilement utilisables en tant que telles. Les données doivent être annotées et étiquetées pour donner à la machine des exemples à partir desquels elle pourra inférer des tendances générales et les étendre à tout nouveau cas. C'est la démarche utilisée, par exemple, dans les algorithmes de reconnaissance d'images, qui doivent arriver à associer des images à des objets (par exemple, reconnaître des lettres et des mots à partir d'un texte manuscrit ou d'un document scanné). Même lorsque des étiquetages ne sont pas nécessaires, comme dans le case des traitements dits « non supervisés » qui laissent l'algorithme détecter (par exemple) des groupes pour en identifier les différences *a posteriori*, les données nécessitent d'être nettoyées, triées, préparées. Ce travail, souvent fait à la main par des prestataires recrutés à travers des plateformes spécialisées comme *Amazon Mechanical Turk*, a un coût non négligeable et nécessite d'importants contrôles de qualité (Casilli *et al*, 2019 ; Tubaro et Casilli, 2019).

Un dernier aspect est le coût élevé des outils de collecte et traitement des données. Pour stocker de grandes bases de données et faire tourner des algorithmes parfois très gourmands en puissance de calcul, des équipements informatiques couteux sont nécessaires. On constate depuis quelques années l'impossibilité pour la recherche publique de concurrencer les grandes multinationales du numérique, plus riches en données ainsi qu'en infrastructures computationnelles. La mise en place et la maintenance d'environnements virtuels d'apprentissage est aussi très couteuse, et génère des inégalités entre fournisseurs. Dans le cas des MOOCs par exemple, l'avantage des grandes universités américaines comme Stanford et MIT est évident, d'autant plus qu'elles peuvent compter sur une audience internationale très large du fait de l'usage de la langue anglaise, qui leur permet de mieux maîtriser leurs coûts fixes (Banerjeee et Duflo, 2014). D'autres établissements moins bien ressourcés optent souvent pour des solutions plus modestes, au prix d'une performance technique plus faible et d'un accès à des bases de données de plus petite taille.

Pour toutes ces raisons, les méthodes plus traditionnelles de collecte de l'information, comme l'enquête, le questionnaire et l'observation (participante ou non), restent essentielles à côté des grandes masses de données numériques. Elles donnent accès aux couches de la population qui sont moins à l'aise avec les instruments numériques, ou qui disposent de conditions limitées d'accès (par exemple, dans des zones rurales mal couvertes par les réseaux des télécommunications). Elles permettent aussi de collecter des données complémentaires, susceptibles de fournir des éléments de contexte, indispensables pour interpréter des données factuelles autrement arides. Elles constituent, enfin, une source précieuse d'information pour une recherche publique qui vise à se faire entendre même si elle n'a pas accès aux données massives dont disposent de grandes entreprises privées. De même, elles peuvent aider les

[4] Règlement Général sur la Protection des Données, entré en vigueur en 2018.

établissements éducatifs qui manquent de l'infrastructure nécessaire pour mettre en place des environnements d'apprentissage en ligne, mais adoptent tout de même une approche *evidence-based*.

Les *big data*, un potentiel encore à découvrir ?

Mais alors, le jeu en vaut-il la chandelle ? Notons d'abord que les remarques ci-dessus visent à démonter une rhétorique qui un peu aveuglement, s'était au début laissée entraîner par l'enthousiasme autour des données numériques. Elle ne vise pas pour autant à en nier les atouts indéniables. Les *big data* nous apportent des informations sur lesquelles d'autres sources restent silencieuses. En sociologie, je peux citer par exemple le travail sur la formation du couple de Bergström (2018), qui montre comment les données d'enquête reflètent le récit d'un couple déjà formé qui revient sur son histoire, alors que les données extraites d'un site de rencontre font voir le couple en train ou même avant de se former – et révèlent ainsi des aspects que la narration *ex post* ne saurait restituer. Dans le domaine de l'éducation, les données numériques extraites des espaces informatiques d'apprentissage restituent une vision très détaillée des parcours des apprenants. Pierrot (2019) documente comment ces données lui ont permis de « mettre en évidence un processus d'appropriation sociale du numérique » qu'il aurait été difficile d'observer autrement. Ces données peuvent également faciliter la mise en place de dispositifs de formation, par exemple en termes de personnalisation des ressources pédagogiques, d'identification des risques de décrochage, ou de meilleure compréhension des effets de différents types de *feedback* (Peraya, 2019a).

Mais donc, comment bénéficier au mieux de ces avantages, tout en minimisant les risques soulignés plus haut ? Les contributions au débat lancé dans la revue DMS m'inspirent trois grandes réponses. La première est qu'il faut intégrer l'usage des *big data* en général, et des LA en particulier, dans une approche éthique. Nombreux sont les participants au début qui ont souligné l'importance de la transparence de la démarche à chacune de ses étapes (de la récolte au traitement et à l'utilisation des données numériques), et du besoin de responsabilisation des auteurs (Gras, 2019 ; Peraya, 2019b). Concrètement, l'adoption d'une approche éthique exige la mise en œuvre d'algorithmes « explicables »[5], de formations et de campagnes de sensibilisation des utilisateurs/apprenants, et d'un cadre de gouvernance garantissant le maintien du contrôle par les utilisateurs, avec la possibilité d'ajuster leur paramétrage et éventuellement de quitter le système sans craintes s'ils le souhaitent. J'ajouterais qu'il s'agit d'aller quelque peu au-delà du simple respect des normes (comme le RGPD) pour rendre l'outil bénéfique pour tous. J'interprète en ce sens le plaidoyer de Boyer (2019) en faveur d'une algorithmique « bienveillante », qui valorise les efforts et l'engagement de l'apprenant, dans l'objectif de l'accompagner et de le soutenir.

Il faut également que les LA s'intègrent dans une approche inter-disciplinaire. Romero (2019) le dit bien : « il est souhaitable de pouvoir encourager la coopération entre chercheurs en sciences de l'éducation et en sciences du numérique ». De cette manière, on peut construire de meilleurs modèles d'analyse de traces d'apprentissage, tout en veillant à la constitution d'un cadre éthique comme il a été esquissé ci-dessus. Les expériences des auteurs ayant participé au débat sont d'ailleurs parlantes, une partie d'entre eux provenant du monde des sciences de l'éducation, les autres de l'informatique. Cette coopération n'est certes pas simple : bien que formellement prisée par nombre d'institutions, et de plus en plus souvent exigée dans les appels d'offre pour la recherche, l'inter-disciplinarité se heurte à une organisation de l'enseignement, de l'évaluation des chercheurs, et de la ligne éditoriale des revues qui sont encore, très largement, mono-disciplinaires. Il faut faire preuve de ténacité et parfois de créativité

[5] Explicabilité et interpretabilité constituent un domaine de recherche fleurissant aujourd'hui en informatique (voir par exemple Escalante *et al*, 2018).

pour sortir d'impasse, mais les contributeurs à ce débat font croire que ce chemin à l'apparence dur peut mener à de très bons résultats.

Le dernier besoin qu'il me semble nécessaire de mettre en avant est un contexte organisationnel adapté, permettant l'implication d'acteurs multiples : des chercheurs issus de disciplines différentes, premièrement des sciences de l'éducation et de l'informatique (mais aussi, peut-être, du droit, de l'éthique, voire de la gestion ou de l'ergonomie), ainsi que des enseignants (*a priori*, dans toute discipline), des techniciens et/ou assistants, des responsables administratifs, et bien sûr des apprenants. C'est seulement par la collaboration entre toutes les parties prenantes qu'il est possible de construire une solution de LA éthique et véritablement utile à l'amélioration des parcours d'apprentissage, tout en produisant les informations nécessaires pour la recherche. Là aussi, le chemin peut être ardu, nécessitant une souplesse institutionnelle qui n'est pas toujours au rendez-vous, mais il faut essayer : Boyer (2019) souligne comment l'implication d'une multiplicité d'acteurs a constitué, dans son expérience, un facteur essentiel pour aborder les LA conjointement sous leurs différents aspects.

Conclusion

En conclusion, émerge de ce débat une vision des *big data* en général, et des LA en particulier, qui sans se laisser emporter par les enthousiasmes du début, reste optimiste quant aux opportunités ouvertes par ces nouvelles données et techniques. La clé est de ne pas les imaginer comme une sorte de mécanisme automatique, qui à lui seul serait capable de transformer nos pratiques d'enseignement et de recherche, en anéantissant au passage tous nos savoirs et outils traditionnels : c'est en réalité la structure humaine qui se forme autour des données, la collaboration et l'interaction, l'accord sur les valeurs et normes à respecter, qui va en déterminer l'utilité et le succès. Le cadrage éthique, la coopération inter-disciplinaire, et l'implication de tous les acteurs, n'ont rien d'automatique et relèvent de l'organisation de nos activités et des relations inter-personnelles. C'est à ce niveau qu'il faut agir pour que l'investissement dans les technologies apporte les fruits espérés.

Références

Anderson, C. (2008). The end of theory: The data deluge makes the scientific method obsolete. *Wired*, 16, 7: https://www.wired.com/2008/06/pb-theory/

Banerjee, A.V. et Duflo, E. (2014). (Dis)organization and success in an economics MOOC. *American Economic Review*, 104 (5), 514-518.

Bastin, G. et Tubaro, P. (2018a). Direction d'un numéro spécial de la *Revue française de sociologie* (n. 59(3)) sur « Big data, société et sciences sociales ».

Bastin, G. et Tubaro, P. (2018b). Le moment *big data* des sciences sociales. *Revue française de sociologie*, 59(3), 375-394.

Bergström, M. (2018). De quoi l'écart d'âge est-il le nombre : L'apport des *big data* à l'étude de la différence d'âge au sein des couples. *Revue française de sociologie*, 59(3), 395-422.

Boyer, A. (2019). Quelques réflexions sur l'exploration des traces d'apprentissage. *Distances et médiations des savoirs* [En ligne], 27, mis en ligne le 13 octobre 2019, URL : http://journals.openedition.org/dms/4086

Casilli, A., Tubaro, P., Le Ludec, C., Coville, M., Besenval, M., Mouhtare, T. et Wahal, E. (2019). *Le micro-travail en France. Derrière l'automatisation, de nouvelles précarités au travail ?* Rapport final du projet *Digital Platform Labor* (DiPLab).

Escalante, H.J., Escalera, S., Guyon, I., Baró, X., Güçlütürk, Y., Güçlü, U., van Gerven, M. (dir., 2018). *Explainable and Interpretable Models in Computer Vision and Machine Learning*. Zurich: Springer, The Springer Series on Challenges in Machine Learning.

Gras, B. (2019). Éthique des *Learning Analytics. Distances et médiations des savoirs* [En ligne], 26, mis en ligne le 17 juin 2019, URL : http://journals.openedition.org/dms/3768

Hargittai, E. (2018). Potential biases in big data: omitted voices on social media. *Social Science Computer Review*, 1-15.

Peraya, D. (2019a). Les *Learning Analytics* en question. *Distances et médiations des savoirs* [En ligne], 25, mis en ligne le 24 mars 2019, URL : http://journals.openedition.org/dms/3485

Peraya, D. (2019b). Les *Learning Analytics* : contraintes méthodologiques et « gouvernance » éthique des données. *Distances et médiations des savoirs* [En ligne], 26, mis en ligne le 24 juin 2019, URL : http://journals.openedition.org/dms/3739

Peraya, D. (2019c). Entre l'enseignement et la recherche, quelle place pour les *Learning Analytics* ? *Distances et médiations des savoirs* [En ligne], 27, mis en ligne le 13 octobre 2019, URL : http://journals.openedition.org/dms/4080

Peraya, D. et Luengo, V. (2019). Les *Learning Analytics* vus par Vanda Luengo. *Distances et médiations des savoirs* [En ligne], 27, mis en ligne le 13 octobre 2019, URL : http://journals.openedition.org/dms/4096

Pierrot, L. (2019). Les LA : des réponses et des promesses. *Distances et médiations des savoirs* [En ligne], 26, mis en ligne le 17 juin 2019, URL : http://journals.openedition.org/dms/3764

Robinson, L., Cotten, S.R., Ono, H., Quan-Haase, A., Mesch, G., Chen, W., Schulz, J., Hale, T.M. et Stern, M.J. (2015). Digital inequalities and why they matter. *Information, Communication & Society*, 18(5), 569-582.

Romero, M. (2019). Analyser les apprentissages à partir des traces. *Distances et médiations des savoirs* [En ligne], 26, mis en ligne le 17 juin 2019, URL : http://journals.openedition.org/dms/3754

Savage M. et Burrows R. (2007). The coming crisis of empirical sociology. *Sociology*, 41(5), 885-899.

Tubaro, P. (2019). La vie privée, un bien commun ? *Regards croisés sur l'économie*, 23(2), 129-137.

Tubaro, P. et Casilli, A.A. (2019). Micro-work, artificial intelligence and the automotive industry. *Journal of Industrial and Business Economics*, 46(3), 333-345.

Tubaro, P., Casilli, A.A. et Sarabi, Y. (2014). *Against the Hypothesis of the "End of Privacy": An Agent-Based Modelling Approach to Social Media*. Zurich: Springer, SpringerBriefs in Digital Spaces.